FACTORES DE RENDIMIENTO EN ESTUDIANTES

Realizado por: Pavel Alarcón Miranda

Introduccion a Ciencia de Datos

Docente: Jaime Alejandro Romero Sierra

INTRODUCCION

El objetivo de este proyecto es identificar los factores clave que influyen en el rendimiento académico de los estudiantes y desarrollar un modelo predictivo para predecir su calificación final. El proyecto emplea análisis de datos, machine learning y visualizaciones interactivas.

El rendimiento académico depende de múltiples factores como el involucramiento parental, la calidad del entorno escolar y los hábitos de estudio. Entender estos elementos permite diseñar estrategias educativas efectivas para mejorar los resultados, personalizar la enseñanza y apoyar a los estudiantes con un desempeño deficiente.

El análisis se basa en un dataset compuesto por registros de estudiantes, incluyendo variables como asistencia, horas de estudio, calidad del maestro, ingreso familiar. Los datos fueron preprocesados para garantizar su calidad y utilidad en el modelo predictivo

PROCESADO DE DATOS

Se comienza con una visión general del DataFrame, lo cual permite identificar diversas inconsistencias en los datos, tales como: valores nulos, filas duplicadas, valores incorrectos, datos atípicos, tipos de variables que no corresponden a sus valores, y columnas sin datos válidos para su análisis.

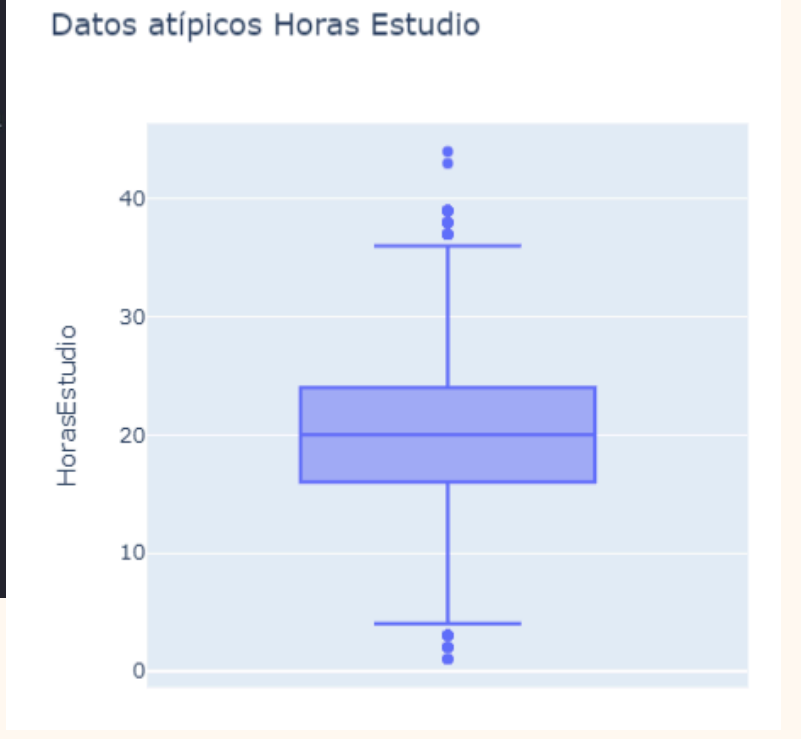
```
#Tipos de datos
df.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6753 entries, 0 to 6752
Data columns (total 20 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Hours_Studied               6687 non-null   object
 1   Attendance                  6703 non-null   object
 2   Parental_Involvement        6695 non-null   object
 3   Access_to_Resources         6696 non-null   object
 4   Extracurricular_Activities  6680 non-null   object
 5   Sleep_Hours                 6696 non-null   object
 6   Previous_Scores             6687 non-null   object
 7   Motivation_Level            0 non-null      float64
 8   Internet_Access             6700 non-null   object
 9   Tutoring_Sessions           6686 non-null   object
 10  Family_Income               6691 non-null   object
 11  Teacher_Quality             6611 non-null   object
 12  School_Type                 6696 non-null   object
 13  Peer_Influence              6687 non-null   object
 14  Physical_Activity           6685 non-null   object
 15  Learning_Disabilities       6684 non-null   object
 16  Parental_Education_Level    6604 non-null   object
 17  Distance_from_Home          6634 non-null   object
 18  Gender                      6686 non-null   object
 19  Exam_Score                  6675 non-null   object
dtypes: float64(1), object(19)
memory usage: 1.0+ MB
```

```
#Cantidad de duplicados
df.duplicated().sum()
```

```
np.int64(54)
```



PROCESADO DE DATOS

A continuación, se procede a realizar diversas tareas de limpieza de datos para garantizar su calidad y utilidad en el análisis. Primero, se eliminan las filas duplicadas, seguidas por la eliminación de la columna “motivation level”, debido a la ausencia de datos utilizables.

Posteriormente, se identifican y eliminan los valores atípicos. Los valores NaN y los datos erróneos en las variables numéricas se reemplazan por el promedio, mientras que, en el caso de las variables categóricas, se opta por eliminar las filas correspondientes.

```
df3.info()
```

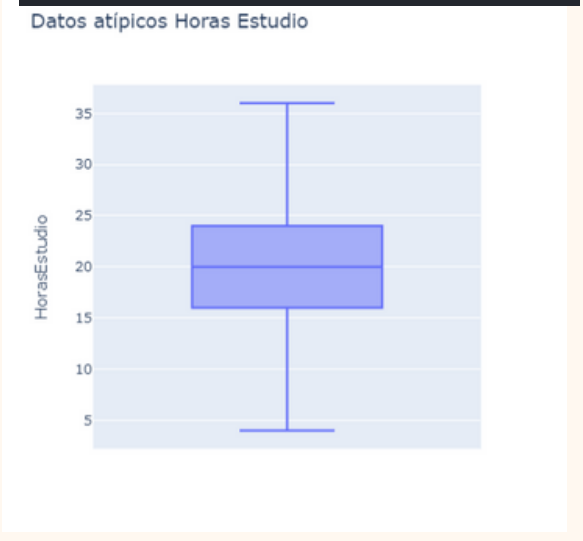
```
<class 'pandas.core.frame.DataFrame'>
Index: 6600 entries, 0 to 6752
Data columns (total 19 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   HorasEstudio                6600 non-null   float64
 1   Asistencia                  6600 non-null   float64
 2   InvolucramientoParental     6600 non-null   object
 3   AccesoRecursos              6600 non-null   object
 4   ActividadesExtracurriculares 6600 non-null  object
 5   Horas de sueño              6600 non-null   float64
 6   ResultadosPrevios           6600 non-null   float64
 7   AccesoInternet              6600 non-null   object
 8   SesionesTutoria             6600 non-null   float64
 9   IngresoFamiliar             6600 non-null   object
 10  CalidadMaestro              6600 non-null   object
 11  TipoEscuela                 6444 non-null   object
 12  InfluenciaCompañeros        6600 non-null   object
 13  ActividadFisica             6600 non-null   float64
 14  ProblemasAprendizaje        6600 non-null   object
 15  NivelEducacionParental      6600 non-null   object
 16  DistanciaACasa              6600 non-null   object
 17  Genero                      6600 non-null   object
 18  PuntajeExamen               6600 non-null   float64
dtypes: float64(7), object(12)
memory usage: 1.0+ MB
```

```
#Elimina los duplicados
df2 = df.drop_duplicates()

df2.duplicated().sum()
```

✓ 0.0s

```
np.int64(0)
```



VISION GENERAL

	HorasEstudio	Asistencia	Horas de sueño	ResultadosPrevios	SesionesTutoria	ActividadFisica	PuntajeExamen
count	4295.000000	4295.000000	4295.000000	4295.000000	4295.000000	4295.000000	4295.000000
mean	20.061001	79.988824	7.044703	67.023749	1.279162	2.964843	67.023749
std	5.718070	11.364883	1.441889	3.161231	0.958171	1.018196	3.161231
min	4.000000	60.000000	4.000000	59.000000	0.000000	0.000000	59.000000
25%	16.000000	70.000000	6.000000	65.000000	1.000000	2.000000	65.000000
50%	20.000000	80.000000	7.000000	67.000000	1.000000	3.000000	67.000000
75%	24.000000	90.000000	8.000000	69.000000	2.000000	4.000000	69.000000
max	36.000000	100.000000	10.000000	75.000000	3.000000	6.000000	75.000000

```
<class 'pandas.core.frame.DataFrame'>
Index: 4295 entries, 0 to 6752
Data columns (total 19 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   HorasEstudio                4295 non-null   float64
 1   Asistencia                  4295 non-null   float64
 2   InvolucramientoParental     4295 non-null   object 
 3   AccesoRecursos              4295 non-null   object 
 4   ActividadesExtracurriculares  4295 non-null   object 
 5   Horas de sueño              4295 non-null   float64
 6   ResultadosPrevios           4295 non-null   float64
 7   AccesoInternet              4295 non-null   object 
 8   SesionesTutoria             4295 non-null   float64
 9   IngresoFamiliar             4295 non-null   object 
 10  CalidadMaestro              4295 non-null   object 
 11  TipoEscuela                 4295 non-null   object 
 12  InfluenciaCompañeros        4295 non-null   object 
 13  ActividadFisica             4295 non-null   float64
 14  ProblemasAprendizaje        4295 non-null   object 
 15  NivelEducacionParental      4295 non-null   object 
 16  DistanciaACasa              4295 non-null   object 
 17  Genero                      4295 non-null   object 
 18  PuntajeExamen               4295 non-null   float64
dtypes: float64(7), object(12)
memory usage: 671.1+ KB
```

VISION GENERAL

El conjunto de datos cuenta con 19 columnas y 4,295 entradas, de las cuales 7 son numéricas, y para estas se calcularon los datos estadísticos correspondientes. Las demás columnas son categóricas.

En cuanto a las horas de estudio, los estudiantes dedicaron en promedio unas 20 horas a estudiar, aunque hubo bastante variación, ya que algunos estudiaron mucho más que otros. En cuanto a la asistencia, la media fue de casi un 80%, lo que muestra que la mayoría de los estudiantes asistieron regularmente a clases.

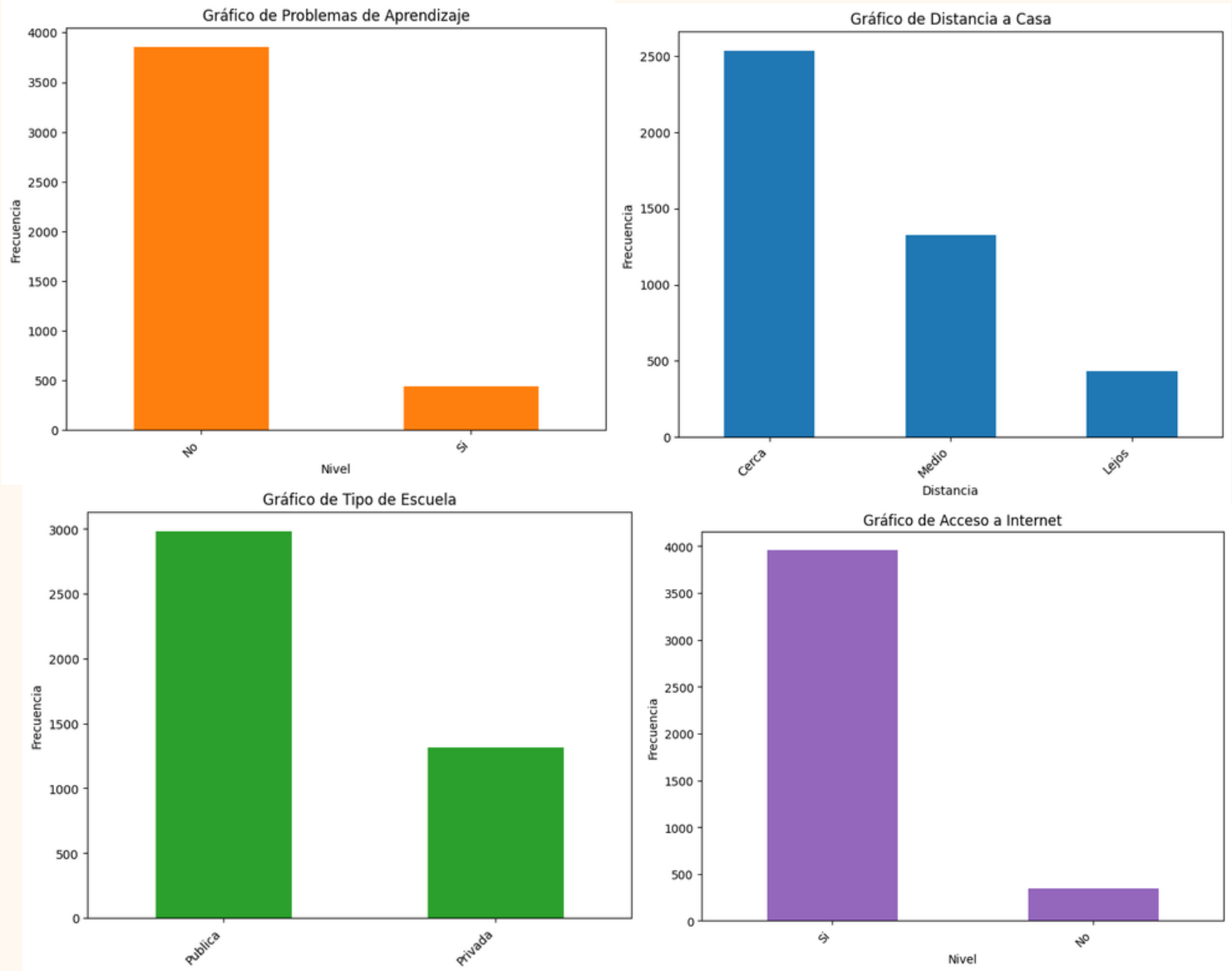
Respecto a las horas de sueño, el promedio fue de alrededor de 7 horas, lo que se considera una cantidad saludable para la mayoría de las personas. Los resultados previos mostraron bastante variabilidad, lo que indica que los estudiantes tenían un rendimiento académico diverso antes del examen final. En cuanto a las sesiones de tutoría, el promedio fue bajo, lo que sugiere que muchos estudiantes no aprovecharon este recurso adicional.

En lo que respecta a la actividad física, la cantidad promedio realizada fue moderada. Finalmente, el puntaje del examen final fue de 67 puntos, con poca variación entre los estudiantes, lo que indica que los resultados fueron bastante similares.

VISUALIZACION

En cuanto a la frecuencia de las categorías, se observa una distribución bastante equilibrada en general, aunque destacan algunos patrones específicos. La mayoría de los estudiantes tiene acceso a internet, y hay aproximadamente el doble de estudiantes en escuelas públicas que en privadas.

Además, se identificó que muy pocos estudiantes presentan problemas de aprendizaje. Por último, la mayoría de los estudiantes vive cerca o a una distancia moderada de la escuela

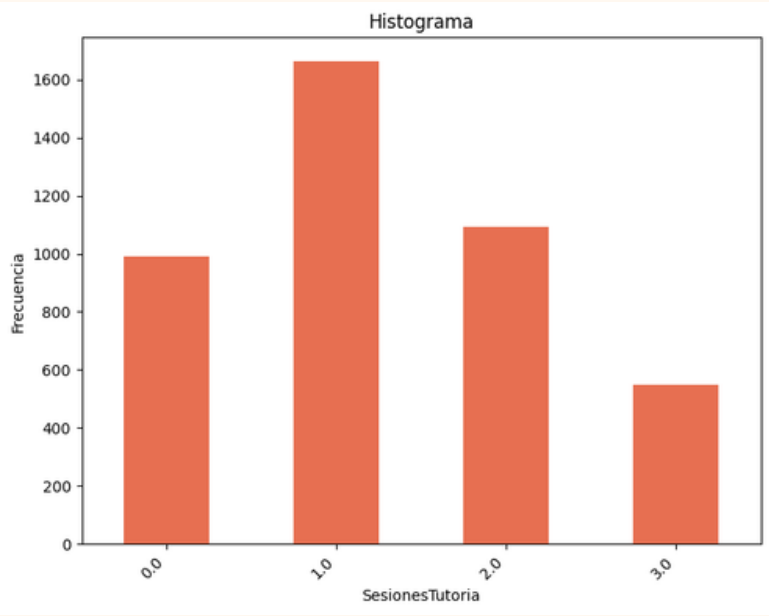


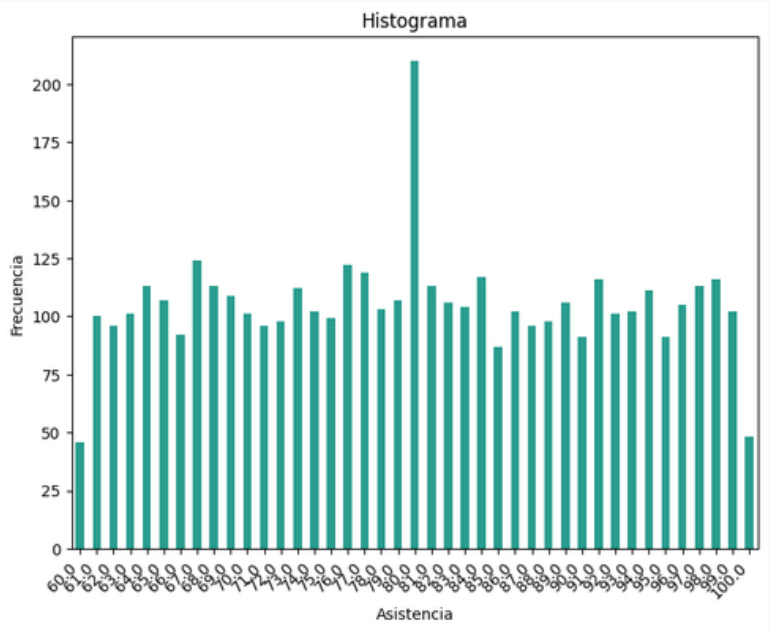
VISUALIZACION

La mayoría de las gráficas muestran distribuciones aproximadamente normales, lo que indica una tendencia equilibrada en los datos.

Sin embargo, hay excepciones notables: el histograma de las sesiones de tutoría presenta una distribución sesgada hacia la izquierda, lo que sugiere que la mayoría de los estudiantes no asisten a estas sesiones.

Por otro lado, el histograma de la asistencia a clases es bastante uniforme, lo que indica que la asistencia se mantiene constante entre los estudiantes.

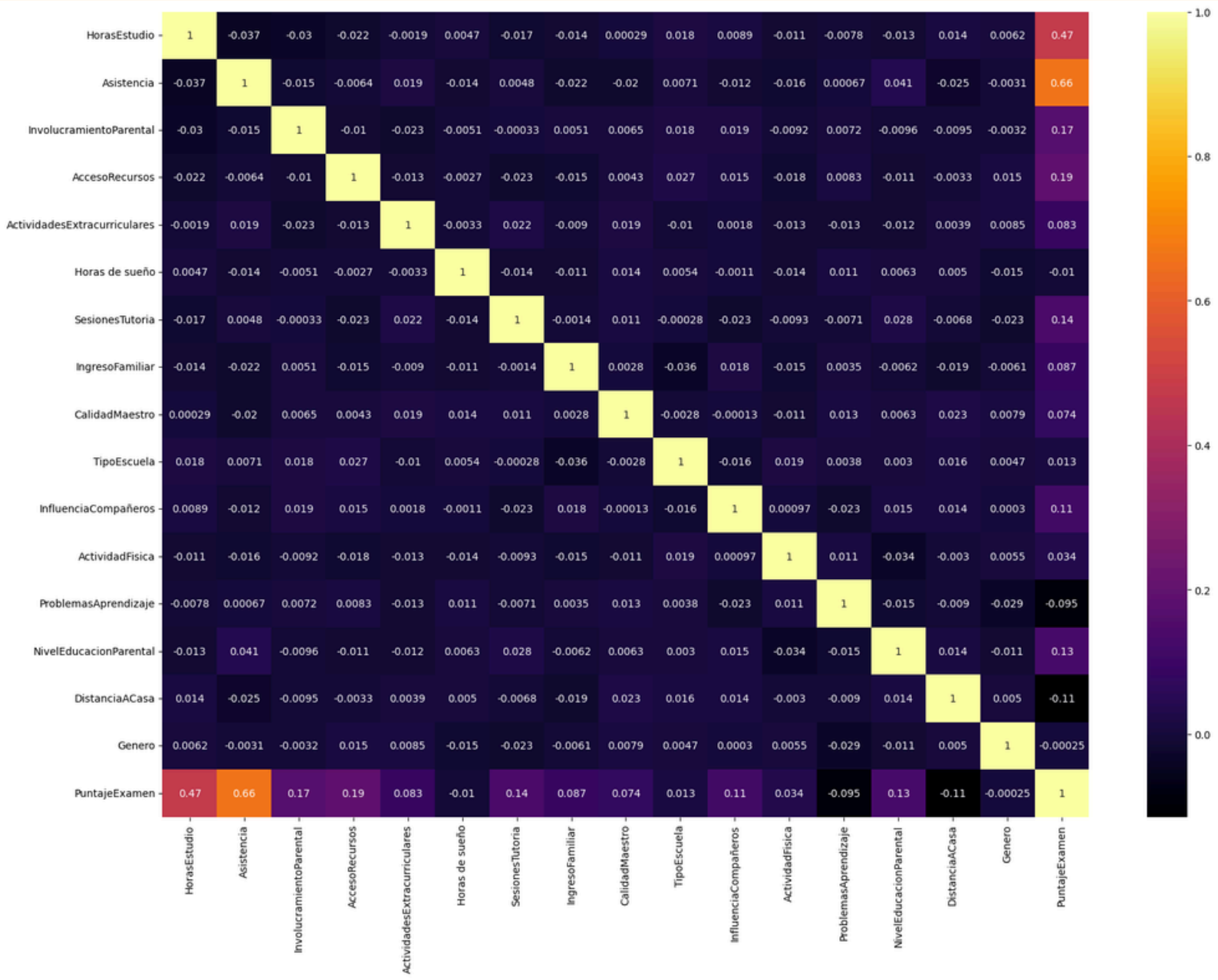




CORRELACION

Al analizar el mapa de calor de correlación, se observa que no existe una relación significativa entre la mayoría de las variables.

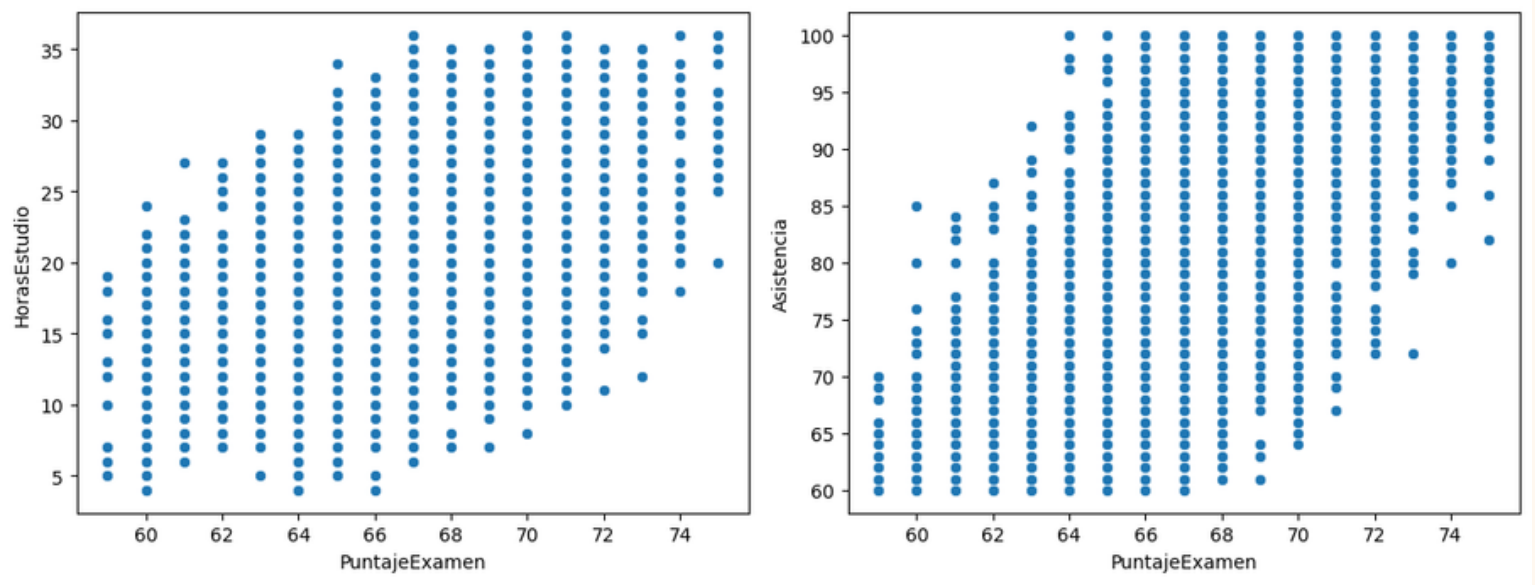
Sin embargo, destacan dos excepciones: el puntaje final muestra una correlación positiva con la asistencia a clases y con el número de sesiones de tutoría, lo que indica que estos factores podrían tener un impacto directo en el desempeño académico de los estudiantes.



CORRELACION

Al analizar más a fondo las relaciones entre las variables, se observa una correlación positiva, pero esta es moderada. Esto indica que existe cierta relación entre las variables, aunque no es lo suficientemente fuerte como para predecir una variable basada en la otra con alta precisión.

Es probable que existan otros factores adicionales que también influyen en los resultados y que deberían ser considerados para obtener un análisis más completo.

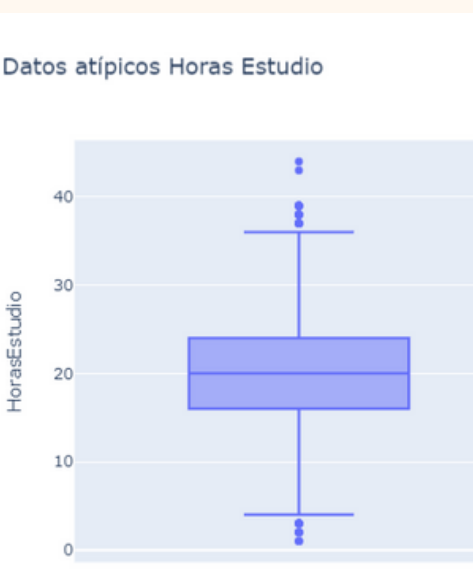


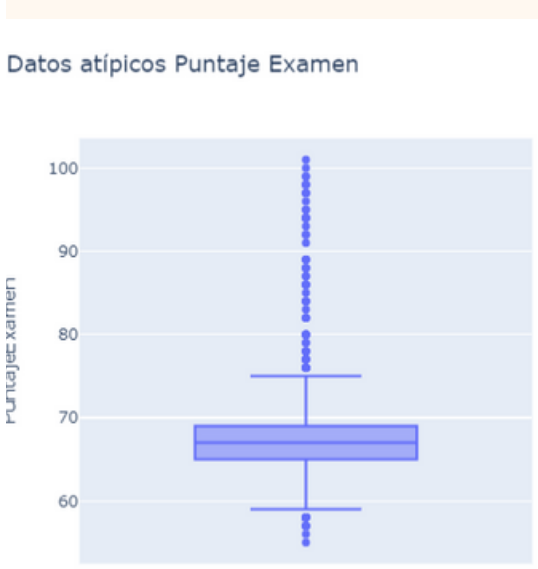
VALORES ATIPICOS

Para identificar valores atípicos, se generaron boxplots para cada variable numérica. A través de esta visualización, se detectaron valores atípicos en las columnas Horas de Estudio, Sesiones de Tutoría y Puntaje del Examen.

Posteriormente, se utilizó el rango intercuartílico para determinar un rango adecuado de análisis. Este método consistió en calcular la diferencia entre el tercer cuartil (Q3, 75%) y el primer cuartil (Q1, 25%), y definir como rango válido los valores entre [Q1−1.5×IQR,Q3+1.5×IQR]

Los valores fuera de este rango fueron considerados atípicos y manejados de acuerdo con los objetivos del análisis. Este enfoque permitió garantizar la calidad de los datos para el modelo predictivo y reducir el impacto de posibles distorsiones.





VALORES FALTANTES

Para identificar los valores faltantes en el conjunto de datos, utilicé el comando isnull(), el cual permitió detectar los valores faltantes en cada columna del DataFrame.

En el caso de las variables numéricas, decidí rellenar los valores faltantes con el promedio de la columna,para mantener la consistencia de los datos sin perder demasiada información.

Para las variables categóricas, opté por eliminarlas debido a que los datos faltantes en estas variables no proporcionaban valor significativo para el análisis y no podían ser imputados de manera efectiva sin introducir sesgos.

```
#Cantidad de valores nulos
df.isnull().sum()
```

✓ 0.0s

Hours_Studied	66
Attendance	50
Parental_Involvement	58
Access_to_Resources	57
Extracurricular_Activities	73
Sleep_Hours	57
Previous_Scores	66
Motivation_Level	6753
Internet_Access	53
Tutoring_Sessions	67
Family_Income	62
Teacher_Quality	142
School_Type	57
Peer_Influence	66
Physical_Activity	68
Learning_Disabilities	69
Parental_Education_Level	149
Distance_from_Home	119
Gender	67
Exam_Score	78
dtype: int64	

```
df2['HorasEstudio'].fillna( round(df2['HorasEstudio'].mean()), inplace=True)
df2['Asistencia'].fillna( round(df2['Asistencia'].mean()), inplace=True)
df2['Horas de sueño'].fillna( round(df2['Horas de sueño'].mean()), inplace=True)
df2['SesionesTutoria'].fillna( round(df2['SesionesTutoria'].mean()), inplace=True)
df2['ActividadFisica'].fillna( round(df2['ActividadFisica'].mean()), inplace=True)
df2['PuntajeExamen'].fillna( round(df2['PuntajeExamen'].mean()), inplace=True)
df2['ResultadosPrevios'].fillna( round(df2['ResultadosPrevios'].mean()), inplace=True)
```

✓ 0.0s

```
df6 = df6.dropna()
df6
```

OBSERVACIONES

Lo que más me llamó la atención fue la relación entre el puntaje de examen, la asistencia y las sesiones de tutoría. Dado que el puntaje de examen era la variable principal de interés, pude observar que, aunque la asistencia y las sesiones de tutoría influyen positivamente en el rendimiento, es el conjunto de todas las variables lo que, en última instancia, determina el puntaje final del alumno.

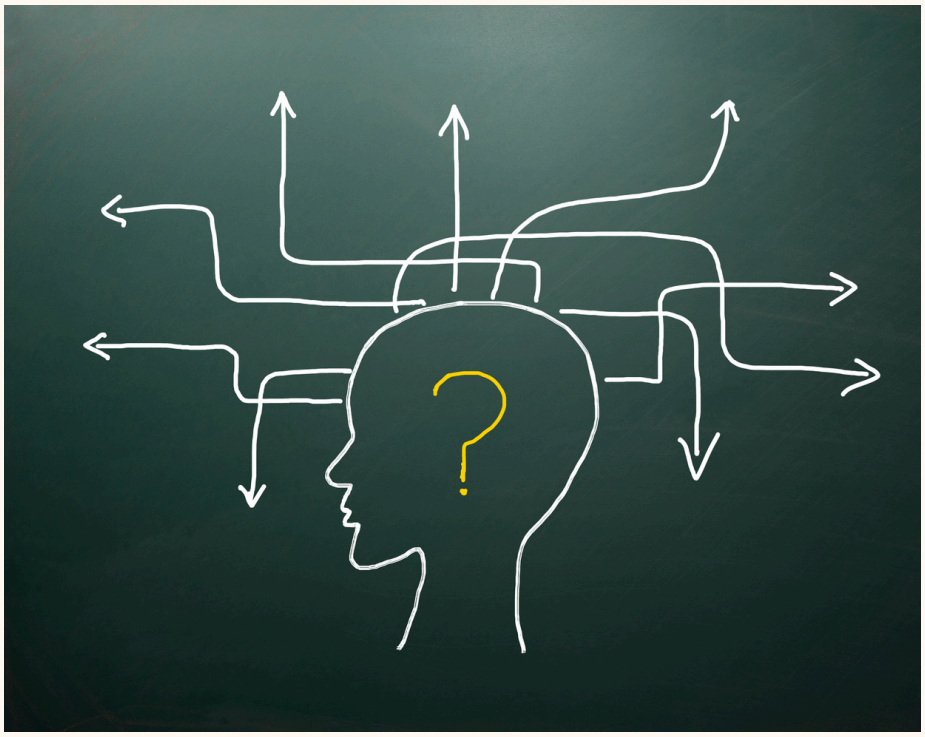
Este hallazgo destaca la importancia de considerar múltiples factores simultáneamente para entender el rendimiento académico de manera integral, en lugar de centrarse únicamente en una variable aislada.

MACHINE LEARNING

El modelo de Machine Learning que utilicé fue un Bosque de Decisiones (Random Forest), ya que contábamos con un gran número de variables, y cada una de ellas influía de manera diferente en la calificación final obtenida por los estudiantes.

Este modelo es particularmente adecuado para manejar conjuntos de datos con múltiples características y relaciones complejas, ya que permite capturar la interacción entre las variables sin necesidad de especificar explícitamente esas interacciones.

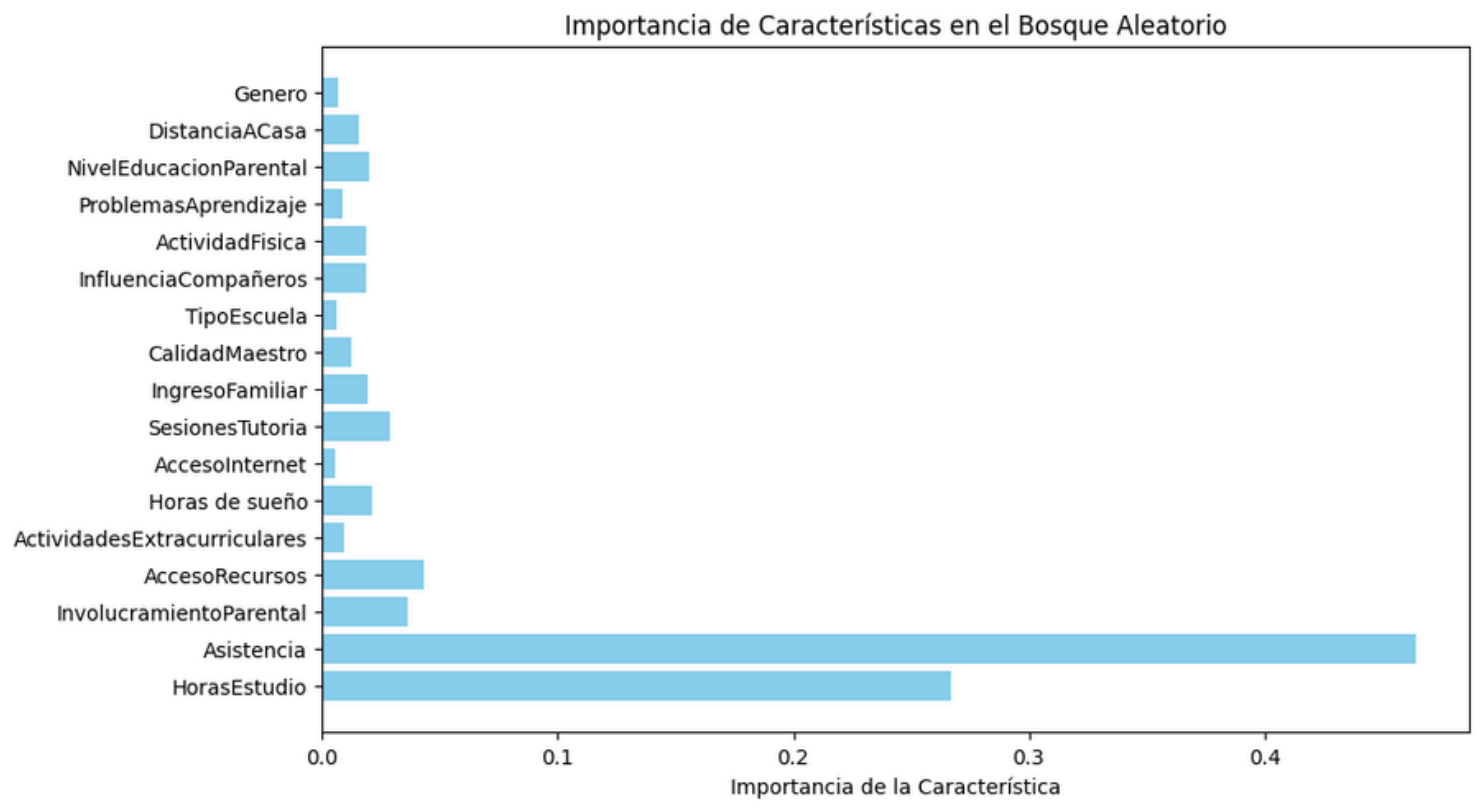
Además, el Bosque de Decisiones es robusto frente a sobreajustes y proporciona una estimación precisa al combinar múltiples árboles de decisión para generar predicciones más confiables.



MACHINE LEARNING

Primero, comencé dividiendo los datos, utilizando el 80% para entrenamiento y el 20% para prueba. Luego, evalué el modelo utilizando el Mean Squared Error (MSE) y el coeficiente de determinación (R^2), obteniendo un 76% de efectividad.

Con este resultado, recurrí a GridSearchCV para encontrar los mejores parámetros para mi modelo. Después de ajustar los parámetros, volví a evaluar el modelo y obtuve un 80% de efectividad. Finalmente, realicé una validación cruzada con 5 particiones, y los resultados de la predicción fueron muy buenos, lo que confirmó la mejora en el rendimiento del modelo.



MACHINE LEARNING

Una vez que el modelo estuvo listo, procedí a realizar una predicción utilizando los factores de rendimiento de un estudiante. Para ello, utilicé las características relevantes del estudiante, como horas de estudio, asistencia, sesiones de tutoría, y otros factores, y alimenté estos datos en el modelo entrenado.

El modelo generó una predicción del puntaje final del estudiante, lo que permitió evaluar cómo las diferentes variables influían en su desempeño académico y obtener una estimación precisa de su calificación final en función de estos factores.

```python
# Características del alumno en un diccionario
alumno = {
    'HorasEstudio': 5,
    'Asistencia': 90,
    'InvolucramientoParental': 3,  # Alto
    'AccesoRecursos': 2,           # Medio
    'ActividadesExtracurriculares': 1,  # Si
    'Horas de sueño': 7,
    'AccesoInternet':1,
    'SesionesTutoria': 2,
    'IngresoFamiliar': 2,          # Medio
    'CalidadMaestro': 3,           # Alto
    'TipoEscuela': 0,              # Pública
    'InfluenciaCompañeros': 1,     # Neutral
    'ActividadFisica': 4,
    'ProblemasAprendizaje': 0,     # No
    'NivelEducacionParental': 2,   # Medio
    'DistanciaACasa': 1,           # Cerca
    'Genero': 0                     # Mujer
}

# Convertir el diccionario a un DataFrame
alumno_df = pd.DataFrame([alumno])

# Predecir la calificación final utilizando el modelo entrenado
prediccion_calificacion = rf_model.predict(alumno_df)

# Mostrar la predicción
print(f"La calificación final predicha para el alumno es: {prediccion_calificacion[0]:.2f}")
```

✓ 0.0s

```
La calificación final predicha para el alumno es: 65.89
```

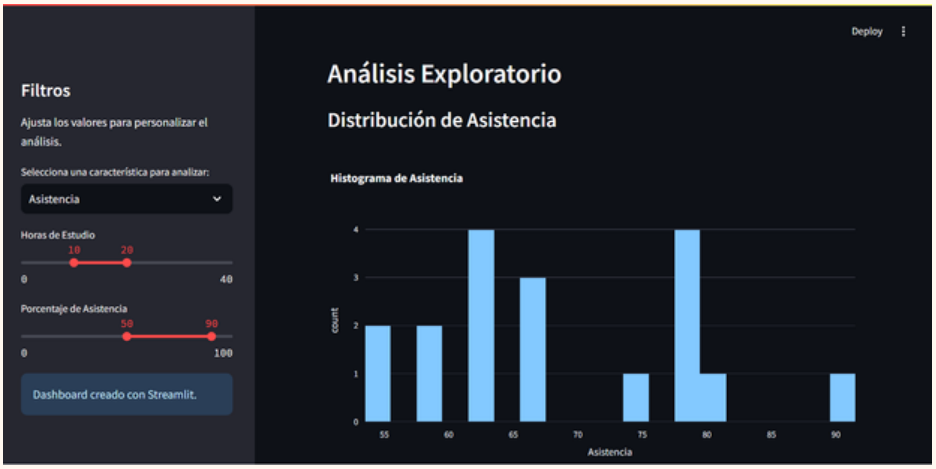
DASHBOARD

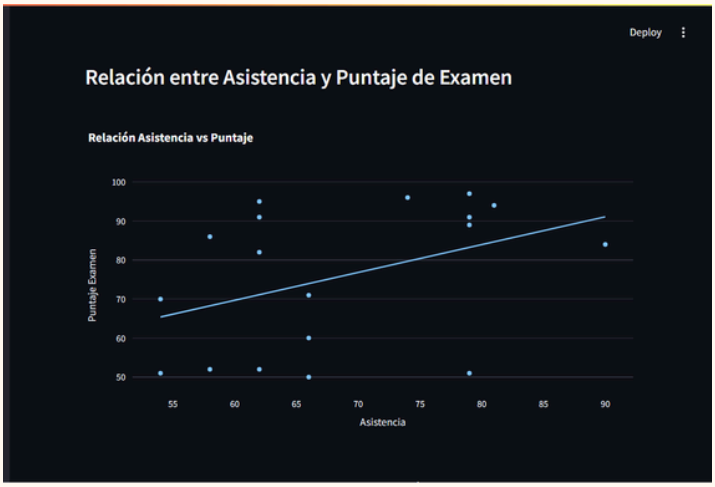
El dashboard tiene como objetivo principal analizar los factores que influyen en el rendimiento académico de los estudiantes, permitiendo realizar predicciones personalizadas basadas en características individuales como horas de estudio, asistencia y sesiones de tutoría.

Entre las visualizaciones incluidas, destacan los histogramas para explorar distribuciones de variables clave, un gráfico de dispersión que muestra la relación entre asistencia y puntaje de examen con tendencias, y un gráfico de barras que identifica las características más influyentes en el modelo predictivo. Además, cuenta con una herramienta que permite predecir el puntaje de un estudiante según sus hábitos y factores, ofreciendo resultados personalizados.

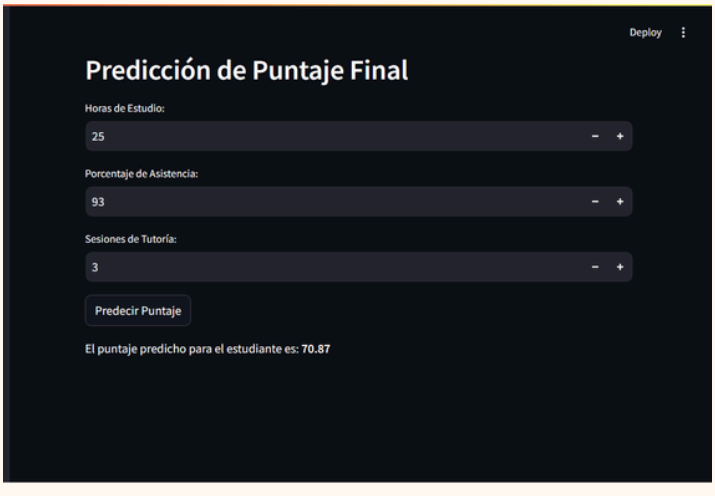
Este dashboard es altamente relevante para la toma de decisiones basada en datos. Ayuda a administradores, docentes y estudiantes a identificar prioridades y áreas de mejora. Proporciona un diagnóstico personalizado que facilita intervenciones específicas para optimizar el rendimiento académico y permite enfocar recursos en los factores más influyentes, como sesiones de tutoría o programas para mejorar la asistencia.

DASHBOARD


Deploy
Filtros
Ajusta los valores para personalizar el análisis.
Selecciona una característica para analizar:
Asistencia
Horas de Estudio
10
20
0
40
Porcentaje de Asistencia
50
90
0
100
Dashboard creado con Streamlit.
Análisis Exploratorio
Distribución de Asistencia
Histograma de Asistencia
count
Asistencia


Deploy
Relación entre Asistencia y Puntaje de Examen
Relación Asistencia vs Puntaje
Puntaje Examen
Asistencia


Deploy
Importancia de las Características en el Modelo
Importancia de Características
Sesiones de Tutoría
Horas de Estudio
Asistencia
Característica
Importancia


Deploy
Predicción de Puntaje Final
Horas de Estudio:
25
Porcentaje de Asistencia:
93
Sesiones de Tutoría:
3
Predecir Puntaje
El puntaje predicho para el estudiante es: 70.87

CONCLUSIONES

El análisis realizado mediante el dashboard ha permitido identificar los factores clave que influyen en el rendimiento académico de los estudiantes, cumpliendo así con los objetivos planteados. Entre los hallazgos principales, se destaca que la asistencia y las sesiones de tutoría tienen un impacto significativo en el puntaje final, como lo evidencia la correlación observada en las visualizaciones y la importancia de estas variables en el modelo predictivo. Asimismo, la herramienta de predicción permite estimar puntajes basados en los hábitos y características de cada estudiante, lo que ofrece un diagnóstico personalizado para orientar intervenciones.

El dashboard ha demostrado ser útil para la toma de decisiones fundamentadas, facilitando la identificación de áreas de mejora tanto a nivel individual como institucional. Por ejemplo, mejorar la asistencia y fomentar la participación en sesiones de tutoría pueden ser estrategias efectivas para incrementar el rendimiento general de los estudiantes.

Recomendaciones

Mejoras en los Datos:

- Ampliar la cantidad de datos recopilados para aumentar la representatividad y reducir posibles sesgos.
- Incluir nuevas variables que puedan influir en el rendimiento, como métodos de enseñanza o factores socioemocionales.
-

Optimización del Modelo:

- Explorar modelos adicionales, como Gradient Boosting o XGBoost, para comparar el rendimiento predictivo con el modelo actual de Random Forest.
- Ajustar hiperparámetros de manera más exhaustiva utilizando validación cruzada con más particiones.
-

Mejoras en las Visualizaciones:

- Incorporar gráficos dinámicos que permitan comparar variables en tiempo real.
- Añadir más detalles explicativos en los gráficos para facilitar su interpretación por parte de los usuarios no técnicos.
- Incluir un resumen ejecutivo visual que consolide los hallazgos clave en una sola vista.

Con estas mejoras, el dashboard podría ser aún más robusto y efectivo, proporcionando insights más precisos y personalizados para mejorar el rendimiento académico en distintos contextos educativos.

REFERENCIAS

Para este análisis, se utilizó la base de datos titulada "Student Performance Factors" disponible en el sitio web Kaggle.

La fuente de la base de datos puede encontrarse en el siguiente enlace: Student Performance Factors - Kaggle.

Además, para el ajuste de los parámetros del modelo de Random Forest, se empleó la técnica de GridSearch.